## Новые возможности поисково-аналитической системы «Галактика-ZOOM»

## (ранжирование документов по значимости)

А.В. Антонов

Kopnopaция «Галактика» alexa@galaktika.ru

Е.С. Курзинер

Kopnopaция «Галактика» koorz@galaktika.ru

**Ключевые слова:** большой текстовый массив, ранжирование, значимость, информационный портрет, скорочтение, тематический контекст.

«Галактика-ZOOM» представляет собой поисково-аналитическую систему для обработки больших текстовых массивов. Эта обработка заключается в получении по запросу так называемого «информационного портрета», или упорядоченного по значимости набора «главных тем» выборки. В основе определения главных тем и их ранга лежит идея о различном распределении слов в конкретной выборке и целой базе. Среди новых возможностей системы — ранжирование документов по значимости, то есть упорядочивание документов внутри выборки по наибольшему соответствию инфопортрета каждого документа инфопортрету выборки в целом. При таком ранжировании наверху оказываются документы, чьи инфопортреты больше похожи на инфопортрет выборки, - наиболее новостная (новая и растиражированная) информация, а инфопортреты «верхних» документов представляют собой своего рода «информационный контекст» запрашиваемой темы. Представляемая технология позволяет в режиме реального времени проводить объективное уточнение и объективную аналитику вопроса.

«Галактика-ZOOM» представляет собой поисково-аналитическую систему для обработки больших текстовых массивов. Эта обработка заключается в получении по запросу так называемого «информационного портрета», или упорядоченного по значимости списка «главных тем» выборки. Выделяемые таким образом значимые слова определяют тематический контекст запрашиваемой информации, что позволяет не только быстро осветить тему, но и эффективно уточнить запрос - даже при поверхностном знании проблематики.

В основе определения главных тем и вычисления их ранга лежит идея о количественном и качественном (лексическом) различии распределения слов в конкретной выборке по отношению к целой базе. Таким образом, главные темы выборки – это слова и словосочетания, «сопутствующие» запросной теме, ассоциируемые с ней в конкретной базе данных.

Среди новых возможностей системы – ранжирование документов по значимости, то есть упорядочивание документов внутри выборки по степени соответствия инфопортрета каждого документа инфопортрету выборки в целом. При таком ранжировании наверху оказываются документы, чьи инфопортреты больше похожи на инфопортрет выборки. Эти документы и будут наиболее релевантными с точки зрения всей выборки – объективно значимыми по отношению к выборке (отражая ее качество, в том числе и возможную «разношерстность») и субъективно значимыми по отношению к запросу (наверху не обязательно самые лучшие документы; «лучшее» достигается при уточнении запроса, сведении выборки к «монотемности»).

Отранжированные по значимости документы упорядочиваются таким образом, что наверху оказывается наиболее новостная, то есть растиражированная, информация.

Инфопортреты же «верхних» документов представляют собой своего рода «информационный контекст» запрашиваемой темы. При этом мы должны помнить, что инфопортрет — это набор главных тем конкретной выборки, полученной по конкретному запросу. Если запрос неоднозначен или вызывает много различных ассоциаций, то так же разношерстно будет выглядеть и инфопортрет.

В чем практический смысл этой опции - ранжирования по значимости? - В возможности вытаскивать наверх документы, которые наиболее полно отражают всю выборку и общий контекст запрашиваемой темы, а значит, в возможности «скорочтения». В отличие от поиска документов «без ранжирования», когда документы упорядочиваются в случайном порядке (то есть в порядке их загрузки в базу), при использовании функции ранжирования можно быстро понять, «о чём пишут», а уточнив запрос, избавившись таким образом от многотемности и многозначности, можно быстро получить наиболее полный («значимыми словами») документ. Соответственно, чем дальше от «верхушки» выборки, тем более «незначимым» (частным, случайным, банальным) будет документ по отношению к запрашиваемой теме. Если тема неоднозначна (в силу омонимии), то наиболее актуальные значения будут наверху: например, на запрос «Анпилов» по газетной базе первые документы - о Викторе Анпилове (политике), а последние – об Андрее Анпилове (барде).

Данная опция – ранжирование - также дает возможность выявлять *тиражируемые* тексты (такие документы при включенном ранжировании ZOOM выстроит друг за другом), что помогает как выявлению «популярной», новостной, информации, так и распознаванию оригинальных и «плагиатизирующих» изданий.

Для примера разберем несколько запросов: два – из газетной базы, один – из литературной.

## 1. Анпилов

Исходный запрос "анпилов и date:<28.11.2002" (по газетной базе) Инфопортрет выборки:

Число документов равно 5730

Статистика слов	
Имя базы	SupBase
Кол-во документов в базе	6797229
Кол-во слов в базе	10865898
Число словомест	6584955
Относит. энтропия ИнфоПортрета	0.82

Абсол.	Энтропия	Абсол.	Энтропия	Коэфф.	
частота	в базе	частота	в папке	значим.	
в базе		в папке			
159710	11	10656	6.9	43.1	КПРФ
6433	7.1	3053	5.2	18.3	РКРП
121186	11	3890	6.4	11.6	МИТИНГ
77429	10	2945	6.6	10.8	ЖИРИНОВСКИЙ
215640	11	4245	6.9	10.7	<b>РИДИКОППО</b>
330481	11	5457	6.1	9.52	ИЗБИРАТЕЛЬ
2135	6.9	993	5.9	9.17	ТЮЛЬКИН
49587	10	2158	6.5	8.72	КИТЧАПМОЖ
146507	9.8	3232	6.6	8.67	ЕЛЬЦИНА
69707	10	2431	6.4	8.54	ЯВЛИНСКИЙ
206907	11	3932	6.2	8.36	СЪЕЗД
421816	9.5	5277	6.7	8.11	ВЫБОРЫ
152192	11	3136	6.1	7.27	ЛЕБЕДЬ
145354	11	2836	6.5	6.99	ЛЕВЫЙ
104937	11	2450	6.1	6.34	ЦК

78148 48765 11353 9096 80415 97021 142880 202450	10 8.3 8.5 11 11 11	1728 1047 959 1880 1956 2345 2684	6.1 6.2 6.5 6.7 6.6	6 5.94 5.59 5.4 5.4	ЭЛЕКТОРАТ  MAKAШОВ  TEPEXOВ  ДЕМОКРАТ  ЛОЗУНГ  ГОЛОСОВАТЬ  ПРЕЛВЫБОРНЫЙ
155864 161297		2486 2314		5.03	БЛОК СТОРОННИК
188662		2613	6.4	1 0 0	прии
11851		918	5.9	4.92	льнин БАБУРИН ИЗБИРАТЕЛЬНЫЙ НПСР
524339	11	4778	6	4.9	ИЗБИРАТЕЛЬНЫЙ
14634	8.3	1137	5.2	4.86	НПСР
22158 9	. 5	1201 6		12.5	КОММУНИСТИЧЕСКАЯ ПАРТИЯ
106731				11.3	
891	6.3				СТАЛИНСКИЙ БЛОК
6571	8.5				RNTYAN RAPOZAY
2865					КОММУНИСТИЧЕСКОЕ ДВИЖЕНИЕ
9445				5.4	
6579	8.2	452	5.3	5.39	ДУХОВНОЕ НАСЛЕДИЕ
55579	11	912	6.2	5.36	СОВЕТСКАЯ ВЛАСТЬ
265977	12			5.21	
2701	7.5	382	4.6	4.9	МОСКОВСКАЯ ОРГАНИЗАЦИЯ
8364	8.7	449	5.3	4.89	ЛЕВАЯ СИЛА
59374	10	899	5.8	4.7	ИЗБИРАТЕЛЬНАЯ КАМПАНИЯ
34655	10	714	5.6	4.55	
53541	10	853	5.6	4.38	ВТОРОЙ ТУР
3330	8	288	5.4	4.15	
24648	7.6	773	4.2	4.04	ИЗБИРАТЕЛЬНОЕ ОБЪЕДИНЕНИЕ
42415	10	671	5.9	3.83	
8219				3.79	
4626	8.2			3.77	
5367	8.3	316	5.2	3.69	КРАСНЫЙ ФЛАГ
7388	8.6	336	5.3	3.59	ЛЕВАЯ ОППОЗИЦИЯ

Из приведенных примеров видно, что Анпилов лишь упоминается в контексте с другими темами (выборы, коммунистическое движение и т.п.), а главная смежная тема – КПРФ, коммунистическая партия. И «верхние» документы, отражающие контекст запросной темы, рассказывают, как Анпилов «вписывается» в этот контекст – его связь с обстоятельствами, движениями, политиками – с которыми он «обручен» в СМИ.

#### Первые документы выборки:

## 1. Московский комсомолец (Москва), N165 (31.8.1999)

… Однако все зависит от того, как пройдут переговоры с ◆Анпиловым ("Трудовая Россия") и Тюлькиным (РКРП). Чистых радикалов в России не так много, и без согласования округов пролететь могут все трое, набрав каждый по 4, 5%, а то и намного меньше. 6. "Трудовая Россия" Численность: 10000. Раньше, когда в РКРП (вторая по численности компартия в России, однако шансы ее на думских

#### 2. Экономическое обозрение Дело, Самара, N 26 (15.7.1997)

... •Анпилова по поводу необходимых перемен во внешности и акценте работников телевидения не оставляют на этот счет никаких сомнений. Равно как и творения А. Проханова и прочих "деньщиков", столь способствовавших "идейному и организационному" росту тов. Зюганова, ныне же претендующих на роль "духовного

центра" лево-патриотической оппозиции. В предвыборной же платформе

#### 3. Сегодня (Москва), N284 (16.12.1999)

... Результат - дамы с **ОВР** не дружат и в третий раз идут на **выборы самостоятельно**. Виктор •Анпилов -Сталинский **блок Бывший** радиожурналист (знает английский и испанский), но **оппоненты** иногда **сравнивают его** с **Шариковым**. **Митинговый оратор**, верный сталинец. В 1993 году поругался с Зюгановым, а тремя годами позже - с **левыми** коммунистами Виктора ...

#### <u>4.</u> <u>МК в Питере (Санкт-Петербург)</u> , N037 (9.9.1999)

… Однако все зависит от того, как пройдут переговоры с •Анпиловым ("Трудовая Россия") и Тюлькиным (РКРП). Чистых радикалов в России не так много, и без согласования округов пролететь могут все трое, набрав каждый по 4, 5%, а то и намного меньше. 6. "Трудовая Россия" Численность: 10000. Раньше, когда в РКРП (вторая по численности компартия в России, однако шансы ее на думских ...

## 5. Советник (Москва), N004 (15.4.1999)

... • Анпилова и "Союза офицеров" С. Терехова), популярны в коммунистической части электората в ряде регионов. Их шансы пройти в Думу по отдельности низки, однако в "отбить" 5% они способны КПРФ свыше избирателей. совокупности Некоммунистические партии протеста Растущее недовольство значительной части общества, по которому в наибольшей степени ударил кризис, могло бы вылиться В поддержку некоммунистических

#### 6. Завтра, N009 (3.3.1998)

... Ее нужно связывать с фигурами Олега Семеновича Шенина, Виктора Аркадьевича Тюлькина, даже Виктора •Анпилова и Альфреда Рубикса,- с теми, кто не считает марксизм-ленинизм досадным наследством. И это даже не драма, а, быть может, трагедия Геннадия Андреевича, что он не может, по большому счету, себя самоидентифицировать. Согласен, что он чувствует себя комфортно в роли лидера парламентской оппозиции. Но

#### 7. Московский комсомолец, N69 (11.4.1996)

... Повторяя слова Виктора •Анпилова о том, что в случае его избрания президентом он первым же указом отменит институт президентства, "законник" Зюганов облек эту идею в более легитимную форму: принять поправки к Конституции об отмене этого института. Все же сказывается думская выучка! Но это красивый пропагандистский жест, не более. Фактически президентский пост - горячо ...

### 8. Час Пик, N020 (27.5.1998)

... Движение распалось на два лагеря - москвич •Анпилов, известный и засвеченный борец за рабочее дело, не поделив власть с питерцем Тюлькиным, был исключен из РКРП и увел за собой часть трудовиков, прихватив регистрационное свидетельство. С РКРП осталась группа "Трудовой России", возглавляемая менее известным депутатом Госдумы Григорьевым из Ленинградской области. Манифест Коммунистической партии в

## 9. Деловые люди (Москва), N65 (30.4.1996)

... По сообщению прессы, сам он считает поведение коммунистов "непорядочным", но ситуацию не комментирует. Лидер "Трудовой России" Виктор •Анпилов также занимает небезинтересную позицию, которую, похоже, разделяет и Станислав Терехов ("Союз офицеров"). В соответствии с ней фигура имеет Геннадия Зюганова не никакого значения. Нο коль реальный претендент УЖ скоро ОН

#### 10. Московский комсомолец, N61 (2.4.1997)

... Значит ли это, что коммунисты смирились со сползанием партии в социал-патриотизм? Понятна горечь одного из делегатов пленума: "В РКРП (Тюлькин, •Анпилов. - А.З.) ушла наиболее активная часть КПСС. Они более красные..." ЧТО СЪЕЗД ГРЯДУЩИЙ ИМ ГОТОВИТ Съезду предшествовали собрания первичек ...

### Последние документы:

## 5729. Афиша (Москва), N019 (14.10.2002)

... 19.00

Перекресток •Анпилов. 19.00 Владимир Бережков, Андрей Петровский гульбарий Александр Софронов. 19.00 19.00 Политехнический музей Хомчик, Сергей Войтенко (баян). Галина Центральный дом ученых Виктор Соломенцев. К 100-летию Вадима Козина. 18.30 ЮМОР Корона Команды КВН "Уездный город" (Магнитогорск), сборная Петербурга. 0.00 КИНО **Культ** "Барака" (Рон Фрике, 1992), "Койаанискацци" (Год-фри Реджио, 1983). 23.00 ДРУГОЕ **MyXA** Открытие выставки Анатолия Брусиловского

#### 5730. Афиша (Москва), N021 (29.10.2001)

... 19.30

Театр "Содружество актеров Таганки" Проект ТС "АЗИЯ+" и Андрей ●Анпилов. 19.00 Центр авторской песни Тимур Шаов. 19.00 \*\*\* ПЯТНИЦА, 9 НОЯБРЯ Большой зал Политехнического музея Жанна Бичевская. "Тихая осень идет". 19.00 "Барды Программа смеются". 18.30 глухаря Концертный зал им. П.И.Чайковского Олег Погудин с программой "Городской романс". 19.00 Театр песни ...

- о совсем другом Анпилове (менее «значимым» в общественной жизни) и совсем другие издания («Афиша») и качество статей (перечень объявлений).

## 2. Гоголь

Запрос: author:Гоголь (по литературной базе)

держи,

дурак!

В первом же документе сочетающиеся мало у кого несочетаемые «радужное» и «черно-белое»: «какая сверкающая, чудная, незнакомая земле даль! Русь!..» «Все сделаешь и все прошибешь на свете копейкой»

очень характерный пример того, что наиболее релевантным запросу оказывается текст, в наибольшей степени отличающий данного автора (Гоголя) от других писателей – поэма о «ненормальной» Руси: поэма «Мертвые души».

#### 1. MERTVYE1.TXT

... ожидания очи?.. И еще, полный недоумения, неподвижно стою я, а уже главу осенило грозное облако, тяжелое грядущими дождями, и онемела мысль пред твоим пространством. Что пророчит сей необъятный простор? Здесь ли, в тебе ли не родиться беспредельной мысли, когда ты сама без конца? Здесь ли не быть богатырю, когда есть место, где развернуться и пройтись ему? И грозно объемлет меня могучее пространство, страшною силою отразясь во глубине моей; неестественной властью осветились мои очи: у! какая сверкающая, чудная, незнакомая земле даль! Русь!.. Держи, Селифану Чичиков

кричал

Вот я тебя палашом! - кричал скакавший навстречу фельдъегерь с усами в аршин. - Не видишь, леший дери твою душу: казенный экипаж! - И, как призрак, исчезнула с громом и пылью ...

## 2. MERTVYE1.TXT

... или просто дурь, только, сколько ни хлестал их кучер, они не двигались и стояли как вкопанные. Участие мужиков возросло до невероятной степени Каждый наперерыв совался с советом: "Ступай, Андрюшка, проведи-ка ты пристяжного, что с правой стороны, а дядя Митяй пусть сядет верхом на коренного! Садись, дядя Митяй!" Сухощавый и длинный дядя Митяй с рыжей бородой взобрался на коренного коня и сделался похожим на деревенскую колокольню, или, лучше, на крючок, которым достают воду в колодцах. Кучер ударил по лошадям, но не тутто было, ничего не пособил дядя Митяй. "Стой, стой! - кричали мужики. - Садись-ка ты, дядя Митяй, на пристяжную, а на коренную пусть сядет дядя Миняй!" Дядя Миняй, широкоплечий мужик с черною, как уголь, бородою и брюхом, похожим на тот исполинский ...

#### Инфопортрет:

#### Число документов равно 184

Статистика слов	
Имя базы	LitBase
Кол-во документов в базе	131254
Кол-во слов в базе	1766947
Число словомест	290985
Относит энтропия ИнфоПортрета	0 9

Абсол.	Энтропия	Абсол.	Энтропия	Коэфф.	
частота	в базе	частота	в папке	значим.	
в базе		в папке			
555			3.7	54.2	ЧИЧИКОВ
	4.4		3.6	36.5	козак
	3.5	269		19.4	ГОРОДНИЧИЙ
11243			3.4	13.2	ИВАНОВИЧ
	4.6		3.3	12.4	БРИЧКА
1374			3.5	10.8	XATA
	5.1		2.9	9.75	TAPAC
4552			3.2	9.71	ПАН
25682			3.1	9.1	ИВАН
	3.6		2.8	8.74	ПАРУБОК
98	3.3	70	2.6	8.31	БУЛЬБА
8380		178	3.8	8.04	ЧИНОВНИК
	2.9	137	2.2	7.64	НОЗДРЕВ
	2.9		2.3	7.6	СОБАКЕВИЧ
7854			3.8	7.52	OKO
1480			3.5	6.26	ПОМЕЩИК
6899	8.3		3.9	5.97	ШАПКА
	2.6	_	2.3	5.87	СЕЛИФАН
3522			3.6	5.73	ИЗВОЛИТЬ
	2.5		2.2	5.73	АНДРИЙ
	5.4		3.1	5.54	ЛЮЛЬКА
3861	8.1		3.9	5.51	НЕОБЫКНОВЕННЫЙ
328	5.6		3.4	5.49	ОБОРОТИТЬСЯ
	2.4		2.1	5.27	БОБЧИНСКИЙ
1230			2.7	4.8	ЖИД
4669	8.3	83		4.72	ПОЗАБЫТЬ
205			2.3	4.58	ПОЧТМЕЙСТЕР
3130	7.6	76	3.6	4.58	TABAK
1246	6.5		3.6	4.54	YPE3
	5.6		2.9	4.43	ХЛОПЕЦ
1592	6.2		2.6	4.43	ФЕДОРОВИЧ
387	5.8		3.3	4.4	ПОВОРОТИТЬ
1077	6.7	51	3.4	4.4	СЮРТУК

					C C
	2.3		2.1	19.8	КУРЕННОЙ АТАМАН
	4.5		2.2	19.2	мертвые души
1629			3	16.5	доврый человек
	3.7		2.4	16	СОБАЧИЙ СЫН
	1.8		1.8	14.6	СТАРЫЙ КОЗАК
492	6.2		2.9	13.8	BCE TOMA
	2.4	9		13.6	СТАРЫЙ ТАРАС
107	4.6		2.6	13.3	СИЕ СЛОВО
41	3.4		1.9	12.7	БОГОУГОДНОЕ ЗАВЕДЕНИЕ
111	4.7		2.4	11.7	никакой образ
235	5.4		2.6	11.5	всякая дрянь
846	6.4	21	2.8	11.1	НАШ ГЕРОЙ
27	3.1	8	1.9	9.96	СВОЯ ЛЮЛЬКА
92	4.5	10	2.2	9.73	все поля
202	4.6	14	2.1	9.61	СТАТСКИЙ СОВЕТНИК
12	2.1	7	1.7	9.54	ВРАЧЕБНАЯ УПРАВА
7	1.7	6	1.6	9.38	ЭТОТ ЧИЧИКОВ
5	1.6	5	1.6	9.26	МОЛОДОЙ КОЗАК
2185	7.6	26	3	9.2	СВОЯ СТОРОНА
1860	7.5	24	2.9	8.99	ТАКОЕ СЛОВО
22	3	7	1.7	8.78	ЧЕРТОВ СЫН
49	3.8	8	1.9	8.6	CBOA XATA
98	4.6	9	2	8.29	непродолжительное время
5735	8.5	38	3.4	8.23	ОДНО СЛОВО
12	2.4	6	1.6	8.16	ВОЛОСТНОЙ ПИСАРЬ
35	3.2	8	1.7	8.03	КОЛЛЕЖСКИЙ РЕГИСТРАТОР
13	2.4	6	1.6	8.03	СВОЯ БРИЧКА
20	2.8	8	1.6	7.93	ТРАКТИРНЫЙ СЛУГА
9	2.2	5	1.6	7.89	ОДНА ХАТА
34	3.4	7	1.7	7.89	вся дворня
74	4.1	9	1.8	7.88	ГУБЕРНСКИЙ ГОРОД
618	6.4	14	2.6	7.81	KAKOE MECTO
10	2.3	5	1.6	7.69	ЗНАЧИТЕЛЬНАЯ МИНА
5	1.3	5	1.3	7.66	ПРАВАЯ МАМЕНЬКА
5	1.3	5	1.3	7.66	ТАКОЙ ГОРОДНИЧИЙ
1961			2.9	7.58	иной раз
11	2.4	5	1.6	7.52	СОВЕРШЕННОЕ НЕДОУМЕНИЕ
7	1.3	7	1.3	7.52	ВСЕ КОЗАКИ
7	1.7	5	1.3	7.09	ЧУБАРЫЙ КОНЬ
2657	7.7	24	2.9	6.98	ДРУГОЙ ДЕНЬ
5814	8.6	35	3.4	6.96	ТОТ УЖ
	5.5		2.2	6.82	ВОСКРЕСНЫЙ ДЕНЬ
	6.2		2.3	6.82	НЕЧИСТАЯ СИЛА
57	3.7		1.7	6.82	НАДВОРНЫЙ СОВЕТНИК
26	3	6	1.6	6.79	КРЕПОСТНОЙ ЧЕЛОВЕК

Последние документы -

из «Ревизора»: про женитьбу, про судьбу, про счастье – наиболее банальная в литературе тема.

### 183. REVISOR.TXT

... правила и солидность в поступках. Марья Антоновна (сквозь слезы). Я, право, маменька, не знала... Анна Андреевна. У тебя вечно какой-то сквозной ветер разгуливает в голове; ты берешь пример с дочерей Ляпкина-Тяпкина. Что тебе глядеть на них? не нужно тебе глядеть на них. Тебе есть примеры другие - перед тобою мать твоя. Вот каким примерам ты должна следовать. Хлестаков (схватывая за руку дочь). Анна Андреевна, не противьтесь нашему благополучию, благословите постоянную любовь! Анна Андреевна (с изумлением). Так вы в

нее?.. Хлестаков. Решите: жизнь или смерть? Анна Андреевна. Ну вот видишь, дура, ну вот видишь: из-за тебя, этакой дряни, гость изволил стоять на коленях; а ты вдруг вбежала как сумасшедшая. Ну вот, право, стоит, чтобы я нарочно отказала: ты недостойна ...

#### 184. REVISOR.TXT

... засадят!.. Знаете ли вы, семь чертей и одна ведьма вам в зубы, что... Анна Андреевна. Ах, боже мой, какие ты, Антоша, слова отпускаешь! Городничий (с неудовольствием). А, не до слов теперь! Знаете ли, что тот самый чиновник, которому вы жаловались, теперь женится на моей дочери? Что? а? что теперь скажете? Теперь я вас... у!.. обманываете народ... Сделаешь подряд с казною, на сто тысяч надуешь ее, поставивши гнилого сукна, да потом пожертвуешь двадцать аршин, да и давай тебе еще награду за это? Да если б знали, так бы тебе... И брюхо сует вперед: он купец, его не тронь. "Мы, говорит, и дворянам не уступим". Да дворянин... ах ты, рожа! - дворянин учится наукам: его хоть и секут в школе, да за дело, чтоб он знал полезное. А ты что? - начинаешь плутнями, тебя хозяин ...

## 3. Запрос:Маккартни

(по газетной базе)

### 3. Курьер (Бруклин), N084 (25.8.2001)

... Можно только надеяться - **Джорджу** 58 лет, и дай-то бог еще увидеть **его** играющим и **поющим**. История

Сэр Пол •Маккартни, кажется, счастлив. Ему хорошо: тяжелые времена миновали, и и гениального угрожает. веселой жизни богатого человека ничто уже не 1998 В году, когда его жена Линда умерла OT рака,

### 4. Курьер (Бруклин), N005 (21.2.2000)

... Закончив долгую беседу с "битлами", только-только вкусившими тогда лишь самые первые плоды успеха и еще не предполагавшими, какая слава обрушится на них уже через пару месяцев, я напоследок спросил у них: "Ребята, а как вы думаете, что будет с вами к года?" первому января двухтысячного Джон Леннон, с самого начала бывший у них главным, ответил первым: "К этому времени я уже давно буду В могиле". Ринго "Я небольшой сказал: бы быть парикмахерской". Старр хотел МОНИВЕОХ

## 5. Итоги (Москва), N033 (21.8.2001)

... Можно только надеяться - **Джорджу** 58 лет, и дай-то бог еще увидеть **его** играющим и **поющим**.

 Маккартни любви история Сэр Пол ●Маккартни, кажется, счастлив. Ему хорошо: тяжелые времена миновали, и веселой жизни богатого И гениального человека ничто уже не угрожает. 1998 году, когда его жена Линда

Как видно, 5-й документ дублирует 3-й — как перепечатку в другом издании тех же авторов, но с другим, более «политкорректным» названием («Великолепная четверка» вместо «Четверо идут по миру»). Не будем особенно над этим задумываться: главное, что совсем другой документ, разъединяющий (или соединяющий?) указанные дубли, также повествует о судьбе всех «битлов» - с которыми, естественно, в mass media и ассоциируется Маккартни.

## Заключение

В заключение хотелось бы подчеркнуть достоинства представляемой технологии в рамках актуальности проблемы поиска информации в большом неструктурированном текстовом массиве. «Галактика-ZOOM» выгодно отличается тем, что поиск информации в базе осуществляется без предварительной ее обработки, без предварительного изучения (в целях наилучшей формулировки) запрашиваемой темы. Ключевые слова в найденной по запросу выборке определяются автоматом, они же служат объективным источником для уточнения запроса, и они же служат критерием значимости документа для ранжирования выборки, так как проблема поиска нужной информации в большом объеме, по существу, оказывается релевантной проблеме ранжирования этой информации.

## Литература

- 1. Антонов А. «Две актуальных задачи Служб Безопасности», Сб. ВИНИТИ №9 2000
- 2. Антонов А. «Большие Информационные Объекты», Сб. ВИНИТИ №4 2001.
- 3. Антонов А. «Информационно-поисковая система Galaktika-ZOOM» с элементами анализа на гипермассивах информации, Сб. ВИНИТИ №8 2001.
- 4. Антонов А. «Проблема Выявления Знаний в пресыщенном информацией мире», BYTE-Россия №2 2002.
- 5. Антонов А., Мешков В. «Современные проблемы поисковых систем и некоторые пути их преодоления», Сер. «Аналитика-Капитал», Москва, 2000.
- 6. В. Ю. Добрынин, И.С. Некрестьянов. "Расширение запросов с помощью вероятностного латентного семантического индексирования", Труды третьей всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции", сентябрь, 2001, 151-155, Петрозаводск, Россия.
- 7. Храмцов П. <Моделирование и анализ работы поисковых систем>. <Открытые системы> 1996.
- 8. Thomas Hofmann. Probabilistic latent semantic indexing. In Proc of SIGIR'99, pp.50-57, 1999.
- 9. Антонов А., Кишинский Ю. «Направления развития информационно-поисковых и аналитических систем», Сб. ВИНИТИ №3 2002.
- 10. Лавреш, Шатских, Лавшук, Антонов. Результаты научных исследований информационных систем электронный офис, документооборот, контроль исполнения поручений, Сер. «Аналитика-Капитал», т.6, Москва, 2002.
- 11. Антонов А. «Проблема Выявления Знаний в пресыщенном информацией мире», ВУТЕ-Россия №2 2002

# New Options of "Galaktika-ZOOM" Search Engine

A.V. Antonov

"Galaktika" Corporation

alexa@galaktika.ru

E. Kurziner

"Galaktika" Corporation koorz@galaktika.ru

Key words: huge text data, information portrait, significant words, significance, ranking, fast reading

"Galaktika-ZOOM" is a Search Engine for analysing huge unstructured textual data. On a query one gets so called "Information Portrait", i.e. a set of key words organized according to the words' significance. The idea of *different* words presentation in a whole base vs. a sample underlies defining the key words and their rank.

Among the Engine's new options there is one of documents ranking by significance, i.e. defining a document's relevance according to its information portrait's likeness to that of the whole sample. Thus the top documents are the most *relevant* as characterizing the *whole* sample (and its quality, often multi-theme) and contain 'circulated' news The top document's Information Portraits present something like "information context" of the query's text.